RNA-Seq workflow

Pipeline development

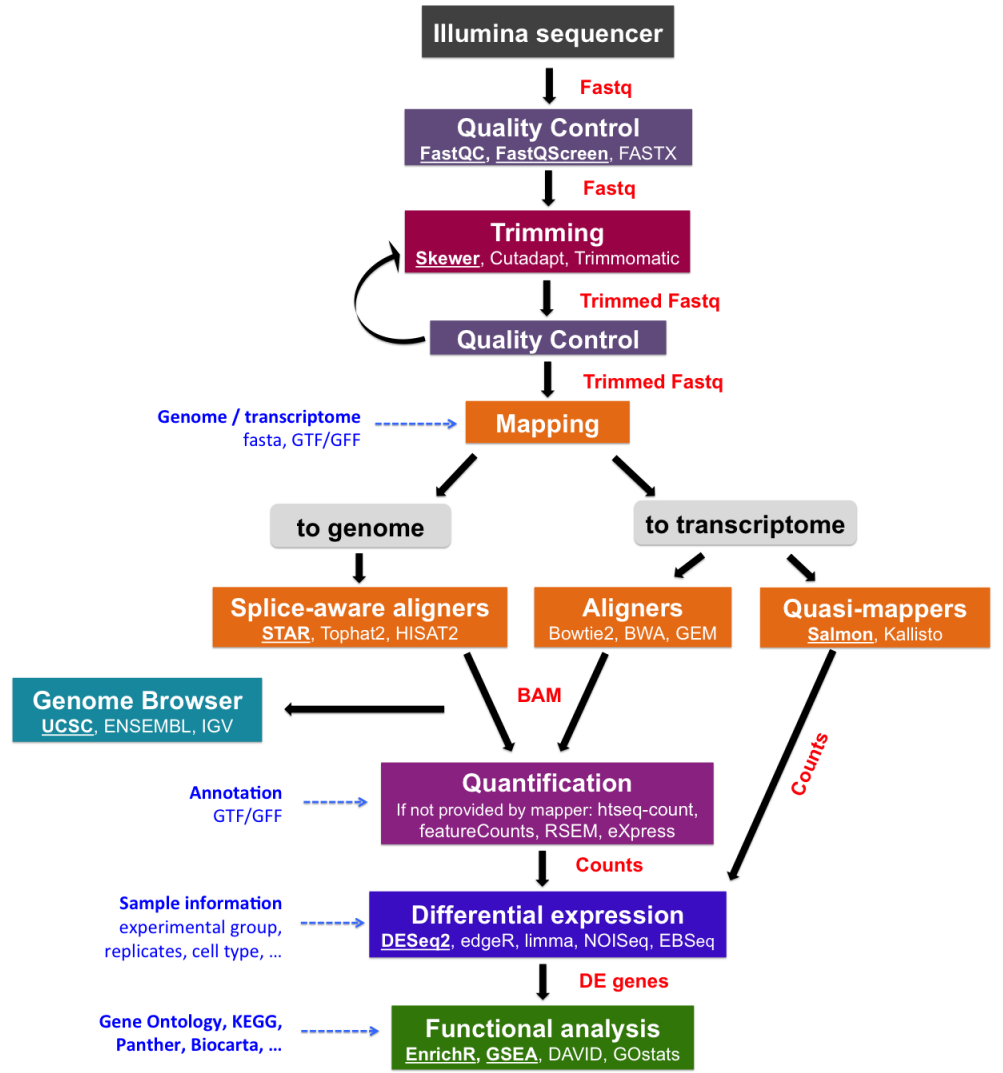
Andrew Ndhlovu

Project SeaStore
Department of Botany and Zoology
Stellenbosch University
https://github.com/PiscatorX/RNA-Seq-devs

July 7, 2023

Bioinformatic pipelines and workflow systems

How do we run (and re-run) analyses in way that saves time?



https://biocorecrg.github.io/RNAseq_course_2019/alignment.html

How do we get started

https://doi.org/10.1371/journal.pcbi.1008645

Building a bioinformatic pipeline

What is a pipeline

A workflow consisting of a variety of steps (commands) and/or tools that process a given set of inputs to create the desired output files.

Syntax	Paradigm	Interaction	Example	Ease of Development	Ease of Use	Performance
Implicit	Convention	CLI	Snakemake, Nextflow, BigDataScript	★★★½	★★★★½	★★★
Explicit	Convention	CLI	Ruffus, bpipe	★★★★	★★★★	★★★
Explicit	Configuration	CLI	Pegasus	★★½	★★★	★★★★
Explicit	Class	CLI	Queue, Toil	★★½	★★★	★★★★★
Implicit	Class	CLI	Luigi	★★★	★★★½	★★★★★
Explicit	Configuration	Open Source Server Workbench	Galaxy, Taverna	★★★	★★★★½	★★★
Explicit	Configuration	Commercial Cloud Workbench	DNAnexus, SevenBridges	★★½	★★★★★	★★★★½
Explicit	Configuration	Open Source Cloud API	Arvados, Agave	★★★	★★★★	★★★★½

https://doi.org/10.1093/bib/bbw020

Most used advanced work flow management systems

snakemake
A framework for reproducible data analysis

COMMON
WORKFLOW
LANGUAGE

Criteria	Snakemake	cwltool	Toil	Nextflow
Installation method used	conda	Python pip	Python pip	conda
Number of RiboViz processing steps implemented	14	3	3	5
Time (person-days) to implement steps	1	1	1[1]	2
Iteration over multiple samples	Yes	See note[2]	See note[2]	Yes
Error recovery strategies	Yes ("keep going" parameter)	See note[2]	See note[2]	Yes (step-specific error strategies)
Aggregation of sample-specific results	Yes	See note[2]	See note[2]	Yes
Conditional execution of steps	Yes (step-specific Python conditions)	No	No[3]	Yes (step-specific "when" clauses)
Step-specific log files	Yes (must be captured explicitly by each step)	See note[2]	See note[2]	Yes (automatically captured)
YAML configuration files	Yes	See note[2]	See note[2]	Yes
"dry run" option	Yes	See note[2]	See note[2]	No (but configuration validation can be implemented)
Output a rerunnable bash script to see the commands that were actually executed	No (outputs a summary file with bash commands)	See note[2]	See note[2]	No (outputs step-specific bash scripts)
Execution within containers, HPC systems, and cloud	Yes	Containers only	Yes	Yes

`https://doi.org/10.1371/journal.pcbi.1008622`

Using a script as pipeline

Save the commands that we run in a file

- Simplest pipeline to implement
- Easiest to learn for beginners
- Usually written in the Unix shell scripting language (Bash)
- Variables and conditional logic can used for flexibility

Simplicity comes at a cost

- Script pipelines (usually) cannot be continued
- Challenging to implement parallel running of tasks
- Become cumbersome and challenging they become longer
- Still requires some understanding of Linux commands and File system

Introduction to Bash shell scripting

Unix shell is both a command-line interface (CLI) and a scripting language, allowing such repetitive tasks to be done automatically and fast

- Many shells e.g. Bourne shell (sh) and C shell (csh)
- Bash (Bourne Again SHell) is the default shell on Linux
- command are saved in text file i.e the script
- On the command line bash is indicated by the prompt

$

- Prompt may be preceded by username, machine name and current directory

Learn more

Lots of tutorials online

- The handout from my supervisor is written for people in biology, it is good start:
  `https://www.bioinf.wits.ac.za/courses/linux/handout.pdf`
- Highly recomment the scripting tutorial on Tutorials point:
  `https://www.tutorialspoint.com/unix/shell_scripting.htm`